

Radiomics approach to quantify shape irregularity from crowd-based qualitative assessment of intracranial aneurysms

Norman Juchler^{a,b*}, Sabine Schilling^{a,c}, Stefan Glüge^a, Philippe Bijlenga^d, Daniel Rüfenachte, Vartan Kurtcuoglu^{b,f,g,h} and Sven Hirsch^a

^aInstitute of Applied Simulation, Zurich University of Applied Sciences, Waedenswil, Switzerland; ^bThe Interface Group, Institute of Physiology, University of Zurich, Zurich, Switzerland; ^cInstitute of Tourism ITW, Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland; ^dNeurosurgery, Clinical Neurosciences Department, University of Geneva, Geneva, Switzerland; ^eDepartment of Neuroradiology, Clinic Hirslanden, Zurich, Switzerland; ^fZurich Center for Integrative Human Physiology, University of Zurich, Zurich, Switzerland; ^gNational Center of Competence in Research, Kidney.CH, Zurich, Switzerland; ^hNeuroscience Center Zurich, University of Zurich, Zurich, Switzerland

Corresponding authors:

Norman Juchler
Institute of Applied Simulation
Zurich University of Applied Sciences
Schloss 1
8820 Waedenswil, Switzerland
Email: norman.juchler@zhaw.ch

Sven Hirsch
Institute of Applied Simulation
Zurich University of Applied Sciences
Schloss 1
8820 Waedenswil, Switzerland
Email: sven.hirsch@zhaw.ch

ORCiDs:

Norman Juchler: [0000-0001-8518-7211](https://orcid.org/0000-0001-8518-7211)
Sabine Schilling: —
Stefan Glüge: [0000-0002-7484-536X](https://orcid.org/0000-0002-7484-536X)
Philippe Bijlenga: [0000-0002-3586-2757](https://orcid.org/0000-0002-3586-2757)
Daniel Rüfenacht: [0000-0001-6538-3186](https://orcid.org/0000-0001-6538-3186)
Vartan Kurtcuoglu: [0000-0003-2665-0995](https://orcid.org/0000-0003-2665-0995)
Sven Hirsch: [0000-0002-5678-4110](https://orcid.org/0000-0002-5678-4110)

Radiomics approach to quantify shape irregularity from crowd-based qualitative assessment of intracranial aneurysms

Abstract. The morphological assessment of anatomical structures is clinically relevant, but often falls short of quantitative or standardized criteria. Whilst human observers are able to assess morphological characteristics qualitatively, the development of robust shape features remains challenging. In this study, we employ psychometric and radiomic methods to develop quantitative models of perceived irregularity of intracranial aneurysms (IAs). First, we collect morphological characteristics (e.g. irregularity, asymmetry) in imaging-derived data and aggregated the data using rank-based analysis. Second, we compute regression models relating quantitative shape features to the aggregated qualitative ratings (ordinal or binary). We apply our method for quantifying perceived shape irregularity to a dataset of 134 IAs using a pool of 179 different shape indices. Ratings given by 39 participants show good agreement with the aggregated ratings (Spearman rank correlation $\rho_{sp} = 0.84$). The best-performing regression model based on quantitative shape features predicts the perceived irregularity with $R^2: 0.84 \pm 0.05$.

Keywords: intracranial aneurysm; morphology; radiomics; multi-rater assessment

Introduction

Linking disease phenotype to image-derived features for computer-aided diagnosis is a central aim in radiomics. While morphological assessment of anatomical structures plays an important role in clinical practice, it is typically based on qualitative, subjective descriptions of phenotypic characteristics. For the clinical use-case of intracranial aneurysm (IA) assessment, we present an approach to translate a qualitative diagnostic judgment of a morphological characteristic into a quantitative metric.

IAs are focal malformations of cerebral arteries, prevalent in 2-5% of the population (Vlak et al. 2011). On average, IAs rupture with an incidence rate of about 1% per year (Wermer et al. 2007). Ruptures lead to haemorrhagic stroke, associated with high mortality and morbidity (Nieuwkamp et al. 2009; Karamanakos et al. 2012). Today, disease status is assessed subjectively, as is the need to treat an aneurysm. An increasing body of literature links irregular aneurysm shape with pathologic wall biology (Frösen et al. 2012; Morel et al. 2018) and increased rate of rupture (UCAS Japan Investigators 2012; Lindgren et al. 2016). Some clinicians have hypothesized this association all along, integrating it into their subjective mental model for making treatment decisions (Ujiie et al. 1999; UCAS Japan Investigators 2012; Kleinloog et al. 2017). “Irregularity” is a vague concept: the medical community has neither developed a common vocabulary to describe irregularity, nor established a standard irregularity rating scheme. As a result, assessments differ between clinicians (Forbes et al. 1996; Suh et al. 2014).

To address these issues, we have developed a method for morphological assessment of IAs that can be generalized to other psychometric quantification problems. Based on data collected with our interactive rating tool for 3D geometries, we show how to aggregate perceived irregularity and judge the degree of consent

(Spearman rank correlation). We compare sub-cohorts of raters (e.g. laypersons vs. clinicians) to assess the test-setup or the inclusion criteria of the raters. Using a pool of geometric shape features, we derive and validate regression models to reproduce the aggregated irregularity ratings. Finally, we break down perceived irregularity into particular morphological attributes (presence of blebs, lobules, rough surface, asymmetry, complex parent vasculature) and again model these quantitatively.

Materials and methods

Given 3D models of the structure under observation, we relate qualitative ratings of morphology to quantitative descriptions of shape through correlation analysis and multivariate regression. In the following, we elucidate our method for the assessment of IA irregularity.

Imaging and patient data

Our dataset comprised 134 saccular IAs (41 ruptured, 78 unruptured, 15 with unknown rupture status) of 110 patients from the University Hospital Geneva (HUG). We extracted geometric models of the aneurysms and the surrounding vasculature from 3D rotational angiographies (3DRAs, voxel sizes in the range of 200-350 μ m) by applying vessel lumen segmentation (geodesic active regions (Bogunović et al. 2010), implemented in the software package GIMIAS (Larrabide et al. 2009)). Standard marching cubes (Lorensen & Cline 1987) was used to convert binary segmentation images into surface meshes. We re-meshed all surfaces using VMTK (Antiga et al. 2008) for a target mesh-cell area of 0.01mm². This step led to more regular meshes and improved overall quality. We assessed the stability of our pipeline regarding different mesh resolutions by comparing the quality of the regression models, exemplarily shown for target mesh area of 0.01 mm² and 0.05mm². (Table A)

Quantitative shape description

We isolated the IA dome with a single planar cut (e.g. (Raghavan et al. 2005)) and computed a set of morphological indices falling into three different types (cf. [Table 1](#)). *Geometry indices* (GIs) capture specific geometric characteristics of the aneurysm dome in a single number. We considered 12 different GIs that primarily measure size or shape (Raghavan et al. 2005; Dhar et al. 2008; Berkowitz 2016). Metrics computed from local

surface properties are termed *distribution-derived indices*. We included features based on *curvature* (both Gaussian and mean curvature, see (Raghavan et al. 2005)) and *surface writhe* (Lauric et al. 2011). Curvature features measure the “bending” and “tortuosity” of the surface, while writhe-based features can be interpreted as a measure of surface asymmetry. Finally, *moment-based descriptors* decompose the surface into different modes. We included Zernike Moment Invariants (ZMI) (Novotni & Klein 2003), which are related to spherical harmonics and compactly represent a 3D surface geometry. Being invariant to scale, translation and rotation of surfaces, ZMIs are well-suited as a basis for comparison of 3D objects. For this study, we used *surface-based* ZMIs (Millán et al. 2007) up to order $n = 20$, resulting in 121 different indices. In total, the pool comprised $d = 179$ different shape indices.

Qualitative shape assessment

The rating tool consisted of two elements: a 3D viewer to examine the object interactively using computer mouse and keyboard, and a rating form to collect the ratings (cf. [Figure 1\(a\)](#)). The written task description emphasized the *qualitative* assessment of shape without providing further clinical information. The raters confirmed having carefully read and understood the instructions before starting the inquiry.

The raters assessed each aneurysm in terms of its shape irregularity on a 9-point rating scale, from “1 – very regular” to “9 – very irregular” (cf. [Figure 1\(b\)](#)). We intentionally refrained from specifying the properties of a perfectly regular/irregular aneurysm. Instead, we relied on the common-sense understanding of geometry and the intuitive nature of the inquiry. To familiarize themselves with the dataset, the participants had to skim through all cases first. After case-by-case assessment with

randomized order, the participants could sort the geometries by increasing irregularity rating and adjust their initial assessment.

We chose a 9-point rating scale to strike a balance between task complexity, rater consistency and informational value: Additional irregularity levels permit a more fine-grained ordering of the cases, but also impair the rater’s ability to consistently sort the cases by increasing irregularity.

As a secondary task, we asked the raters to decide whether the aneurysm under examination exhibited one of the following five morphological attributes: a rough (non-smooth) surface, blebs or lobules, an asymmetric appearance, a complex configuration of the parent vasculature/bifurcation, or none of those (cf. [Table 2](#)). We refer to this part of the inquiry as the (binary) *assessment of morphological attributes*.

A cohort of 39 participants was recruited for the inquiry, which all passed an outlier test (see next section). For each participant, the inquiry resulted in a rough ordering of the cases by perceived irregularity, measured in 9 levels. A subset of 26 raters additionally provided assessments for morphological attributes.

Processing of the rating data

Ordinal rating of irregularity

The varying shape of the rating distributions for each rater (cf. [Figure 2\(a\)](#)) reflects rater subjectivity. To correct for this effect, we ranked the ordinal ratings per rater, where the average ordinal rank for ratings of equal value (tied rank) was computed.

Next, we aggregated the ranked irregularity ratings by computing their means per case. The rating aggregates r_i for case i take values in the range $[1, n]$ where $n = 134$ is the sample size. To normalize this range, we mapped the rating aggregates r_i linearly onto $r'_i \in [0,1]$, with 0 and 1 standing for “very regular” and “very irregular”,

respectively. Hereinafter, we will refer to these normalized, rater-bias adjusted aggregates r'_i as *perceived irregularities*. As a measure of collective agreement, we computed the Spearman rank correlation ρ_{Sp} between perceived irregularities r'_i and the original rating ranks of every rater. To characterize the rater cohort and to test for potential problems with the rating acquisition, we analysed the contribution $\epsilon_j = \sigma_j^2 / \sigma_{tot}^2$ of each rater j to the overall variance

$$\sigma_{tot}^2 = \sum_{i=1}^n \frac{1}{m-1} \sum_{j=1}^m (r'_{ij} - \mu(r'_{ij}))^2 \quad (1)$$

in the data (m : number of raters, n : number of cases, r'_{ij} : normalized rank for rating i of rater j). We applied a robust z-score analysis on the ϵ_j following (Iglewicz & Hoaglin 1993). A rater j was defined to be an outlier if the modified z-score

$$z_{mod}(\epsilon_j) = (\epsilon_j - \tilde{\epsilon}) / \tilde{\sigma}_\epsilon = 0.6745 \cdot (\epsilon_j - \tilde{\epsilon}) / \text{MAD}_\epsilon \quad (2)$$

was larger than 4.0, where $\tilde{\sigma}_\epsilon$ represents a robust estimator for the standard deviation of the ϵ_j , $\tilde{\epsilon}$ and MAD_ϵ denote the median and the median absolute deviation of ϵ_j , respectively, and 0.6745 is the 75th percentile of the standard normal distribution.

Binary ratings of morphological attributes

For each case i and morphological attribute k , we computed the relative counts q'_{ik} of votes in favour of that attribute, normalized by the number of raters. Similar to perceived irregularity, this metric captures how strongly the rater cohort agrees in recognizing a particular morphological attribute. Note that the aggregates q'_{ik} have similar properties to the perceived irregularities r'_i and therefore can be used interchangeably in the subsequent analysis.

Like in the case of perceived irregularity, we also assessed the collective agreement for the ratings of morphological attributes. We considered two methods to

assess the average rater agreement for *binary* ratings of morphological attributes. Fleiss' kappa κ_F measures the agreement within the entire rater cohort, which we evaluated for each morphological attribute separately. Because this first approach ignores any rater-dependent subjectivity, we adopted a second approach in which we compare the binary ratings q_{ijk} of rater j and attribute k (for all cases i) with the binarized aggregates $q_{ik}^{\text{bin},j} = \text{sign}(q'_{ik} - \tau_{jk}^*)$. The binarization threshold τ_{jk}^* is computed for each rater and attribute such that Cohen's kappa κ_C (a measure for inter-rater agreement) between rater j and “binarized average rater” is maximal. In this context, τ_{jk}^* can be interpreted as a perceptual threshold for a rater j to accept the presence of a particular attribute k . [Table 4](#) summarizes the average κ_C and τ_{jk}^* for all 26 raters. Both κ_C and τ_{jk}^* can be used to identify outlier raters using a similar procedure as described in the main article. No such outliers were found in our data.

Association of qualitative ratings and quantitative features

We performed a multivariate analysis to identify “crowd-sourced” shape models that capture perceived morphological characteristics. The size of the feature pool was first reduced by several means: Either we applied principal component analysis (PCA) to identify directions in the feature space with maximal information content, or we ranked and selected relevant features based on univariate linear metrics (correlation coefficients between features and perceived characteristics) or *feature importance*. Feature importance is a statistical measure of how relevant a predictor was in training a potentially nonlinear relationship between the predictor variables (shape features) and response (ratings) with decision trees. To estimate feature importance and to compute non-linear regression, we made use of gradient boosting machines (GBM) provided

through the LightGBM framework (Ke et al. 2017).

Next, we computed multivariate regression models for four different configurations (cf. [Table 3](#)). \mathcal{F}_{univ} represents the set of best performing features from the univariate analysis, \mathcal{F}_{imp} signifies the set of most important features (“importance” as defined above), accounting for 80% of the total importance. For the PCA, the d^* principal components in the (ranked) data space are used, where $d^* < d = 179$ is the number of features that preserve 90% of the overall variance in the data. Instead of ordinary least squares (OLS) regression, we relied on support vector regression (SVR), which is more robust and performed better on our data for higher dimensional feature spaces.

We trained and validated the multivariate models with 5-fold cross-validation and $q = 50$ repetitions. The average root-mean-square error (RMSE) and the coefficient of determination (R^2), computed over the q repetitions, were used as performance metrics to compare the different regression models.

Results

Rating data

We acquired rating data for perceived irregularity of 39 raters from Japan, USA and Europe, all of which passed the outlier test base on the robust z-score. This resulted in a pairwise Spearman rank correlation $\rho_{Sp} = 0.84$ ($p < 0.001$), where ρ_{Sp} was computed between perceived irregularities r'_i and the original ratings, ranked per rater, r'_{ij} .

We also compared the ratings of rater sub-cohorts stratified by professional background. While *clinical experts* rated morphological irregularity on average by 0.467 rating points higher than the *instructed laypersons* (the difference is significant, paired-sample t -test, $p < 0.001$), the resulting *rank*-based aggregate for perceived irregularity cannot be discriminated statistically (paired-sample t -test, $p = 0.967$). As a consequence, the perceived irregularity r'_i is very similar for experts and laypersons, as seen in [Figure 2\(c\)](#).

The level of agreement *per case* i , measured here as the standard deviation σ'_i of (per-rater) ranked irregularity ratings r'_{ij} , varied across cases. A low standard deviation implies a good interrater agreement. σ'_i ranged between 0.050 and 0.261, with a mean of 0.152 (measured in the scale of perceived irregularity $r'_i \in [0,1]$). The agreement was higher between experts than between laypersons ($\bar{\sigma}'_{i,exp} = 0.146$, vs. $\bar{\sigma}'_{i,lay} = 0.151$), but the difference did not reach statistical significance (paired t -test, $p = 0.16$). The best agreement among the raters was observed for extreme cases; very regularly or very irregularly shaped aneurysms were rated the most consistently (cf. [Figure 3](#)).

[Figure 4](#) shows the aggregates q'_{ik} for the morphological attributes in relation to the perceived irregularity r'_i . Interpolation curves (locally weighted scatterplot smoothing, LOWESS) reveal that perceived irregularity is associated with perceived

presence of asymmetry, blebs and lobulations. This trend, however, was not distinguishable for rough surface and complex vasculature.

Multivariate quantitative model for perceived irregularity

Given the rating aggregates (explained variable) and the pool of shape descriptors (predictor variables), we trained statistical models that map feature vectors to ratings. We devised four model configurations (A1, A2, B, C, [Table 3](#)), for which we report the $\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\tilde{r}_i' - r_i')^2}$ as performance metric (cf. [Table 5](#)). RMSE measures the average difference between predicted \tilde{r}_i' and measured perceived irregularity r_i' . We also report the coefficient of determination R^2 , which measures the proportion of the total variance (in the predicted variable) explained by the model. For reference, we give the best-performing *univariate* model, based on the curvature-metric area-normalized L_2 -norm of Gaussian curvature, also known as GLN (Raghavan et al. 2005). This model was trained with the same cross-validation setup for both ranked and metric, non-ranked data. Generally, the inclusion of additional predictors reduced the RMSE score. On ranked data, the prediction error was diminished by about 11% on average, and by about 28% on the metric data. The models predicting the aggregated assessments of morphological attributes generally resulted in a lower prediction performance.

Discussion

In this study, we have collected and aggregated qualitative, ordinal and binary ratings for aneurysm shape. For instance, the perceived irregularity r'_i reflects the collective opinion on the morphological irregularity. The single irregularity ratings per case can vary strongly between participants (cf. [Figure 3](#)), but *rank*-based analysis (Spearman correlation $\rho_{sp} = 0.84$) suggests that raters agree, on average, with the *ordering* of the cases.

This result is robust to local permutations in the ordering of the cases or the exclusion of some raters. Using the aggregated metric allows correction for the inherent subjectivity that comes with irregularity ratings. The results from the subsequent analysis are thus equally robust by design.

The pronounced spread of the ratings around the average is a consequence of the open task formulation, the inconsistency typical of subjective assessment (*intra*-rater disagreement), and the heterogeneous composition of the rater cohort (*inter*-rater disagreement). However, our rank-based method deals robustly with the amount of rater variability.

The level of agreement varies considerably between different cases: extreme cases (very regular, very irregular) are rated more consistently than cases in-between. This variability would ideally be addressed with quantitative criteria to evaluate morphological irregularity. To determine how professional qualification affects ratings, we also compared sub-cohorts of participants. Our results suggest that clinical experience did not affect the judgment of perceived irregularity.

Finally, we developed statistical models to predict perceived irregularity. Such models map quantitative morphological metrics to the subjective assessment of shape, a task that can be considered cognitively complex, involving intuition, experience and

conscious thinking. So far, no quantitative metric exists that specifically measures irregularity of aneurysm shape. A tool to quantify irregularity will help clinicians to assess aneurysms while removing rater subjectivity.

A combination of multiple shape features performed better than univariate models to predict perceived irregularity (cf. [Table 5](#)). A larger model uncertainty (standard deviation of RMSE, [Table 5](#)), as a result of an increased number of model predictors, is overcompensated by increased prediction accuracy. In the case of ranked and metric data, the RMSE improved by 11% and 28%, respectively.

We repeated the analysis for other morphological characteristics for which it is equally difficult to specify robust, quantitative rules. The prediction performances of these models for the aggregates q'_{ik} , however, are poorer. This might be partially explained by the binary assessments carrying less information than ordinal ratings. Binary rating data leads to graded aggregates q'_{ik} , with repercussions on the prediction metrics. Furthermore, the shape features included only insufficiently describe the IA attributes. The development of specific features for these attributes was outside the scope of this study. Regardless of the lower prediction power, we demonstrated that the method can also be applied to binary rating data.

The morphological assessment of anatomical structures is not only relevant for IAs. More generally, the morphology of tissue, bones, organs or vessels, plays an important role in the management of various diseases. We argue that the proposed methodology to capture, normalize and inspect the collective opinion of a rater cohort is equally applicable in other clinical contexts as well. There are two principal requirements for our methodology: 1) The morphology must be assessable by visual inspection, either from 3D surface geometries as in our case, or from 2D or 3D intensity images. 2) A set of quantitative metrics must be computable from the input data (feature

pool) that are thought to capture the qualitative metric (e.g. asymmetry, irregularity, tortuosity).

When working with morphological metrics derived from imaging data, we recommend examining their mesh and resolution dependency. In our use-case, the reduction of mesh resolution (we assessed two surface meshes with average cell areas of 0.01mm² and 0.05mm²) did have a small but noticeable effect on single features (Table A of the supplemental material section). While most metrics are unaffected, curvature metrics are sensitive to mesh resolution. The lower mesh resolution of 0.05mm² yields slightly better correlation coefficients. Fine tuning the mesh size in respect to the imaging resolution holds potential to incrementally improve the model performance.

Putatively, the model accuracies will further increase with a higher number of raters and cases. Although we consider our dataset well-balanced in terms of morphological attributes, it is possible that some characteristics are over- or underrepresented. The features available in our pool might therefore not encompass all morphological attributes that raters take into account, and it is conceivable that metrics exist that encode perceived irregularity more efficiently than the ones we used. We disregard other factors that may have an influence on the morphology of the structure under assessment. In the use-case of IAs presented, for example, a stratification of the aneurysms by location would be an interesting aspect for a follow-up study.

Conclusions

We successfully applied our method to the assessment of IA morphology, for which we trained novel quantitative models for irregularity using qualitative assessments of shape. The inspection of qualitative morphological assessment across multiple raters offers possibilities i) to develop new consensus-based rating-schemes, and ii) to design quantitative tools for the judgement of morphological characteristics. Since the elements of our method do not depend on the particular use-case, our methodology can be useful for the assessment of anatomical structures other than aneurysms.

Acknowledgements

We would like to thank Diana Sapina for her support in preparing the database of intracranial aneurysms, and Victor Garcia and John Bennett for critically proofreading the manuscript.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Raw 3D-DRA of the AneuX test data set were provided by the University Hospital of Geneva and collected with formal patient consent according to the @neurIST protocol and ethics authorization PB_2018-00073 (previously CER 07-05) released June 1st 2007 and renewed April 13th 2010, August 19th 2014 and February 28th 2018 initially by the Geneva Cantonal Ethics Commission for Research involving Humans and renewed by swissethics in 2018.

Disclosure statement

The authors declare that they have no conflict of interest.

Funding

This work was supported by SystemsX.ch, the Swiss initiative in systems biology, under Grant MRD 2014/261 (AneuX project); and by the Swiss National Science Foundation under Grant 147193 (NCCR Kidney.CH).

References

- Antiga L, Piccinelli M, Botti L, Ene-Iordache B, Remuzzi A, Steinman DA. 2008. An image-based modeling framework for patient-specific computational hemodynamics. *Med Biol Eng Comput.* 46:1097–1112.
- Berkowitz BM. 2016. Development of metrics to describe cerebral aneurysm morphology [Internet]. [place unknown]: University of Iowa. Available from: <http://ir.uiowa.edu/etd/2181/>
- Bogunović H, Pozo JM, Villa-Uriol MC, Majoie CBLM, van den Berg R, Gratama van Andel HAF, Macho JM, Blasco J, San Román L, Frangi AF. 2010. Automated segmentation of cerebral vasculature with aneurysms in 3DRA and TOF-MRA using geodesic active regions: An evaluation study. *Med Phys* [Internet]. 38:210–222. Available from: <http://doi.wiley.com/10.1118/1.3515749>
- Dhar S, Tremmel M, Mocco J, Kim M, Yamamoto J, Siddiqui AH, Hopkins LN, Meng H. 2008. Morphology parameters for intracranial aneurysm rupture risk assessment. *Neurosurgery.* 63:185–196.
- Forbes G, Fox AJ, Huston III J, Wiebers DO, Torner J. 1996. Interobserver Variability in Angiographic Measurement and Morphologic Characterization of Intracranial Aneurysms: A Report from the International Study of Unruptured Intracranial Aneurysms. [place unknown]; [cited 2019 Jan 23]. Available from: <http://www.ajnr.org/content/ajnr/17/8/1407.full.pdf>
- Frösen J, Tulamo R, Paetau A, Laaksamo E, Korja M, Laakso A, Niemelä M, Hernesniemi J. 2012. Saccular intracranial aneurysm: Pathology and mechanisms. *Acta Neuropathol.* 123:773–786.
- Iglewicz B, Hoaglin D. 1993. Volume 16: How to Detect and Handle Outliers. In: *ASQC Basic Ref Qual Control Stat Tech.* Vol. 16. [place unknown].
- Karamanakos PN, Von Und Zu Fraunberg M, Bendel S, Huttunen T, Kurki M, Hernesniemi J, Ronkainen A, Rinne J, Jaaskelainen JE, Koivisto T. 2012. Risk factors for three phases of 12-month mortality in 1657 patients from a defined population after acute aneurysmal subarachnoid hemorrhage. *World Neurosurg* [Internet]. 78:631–639. Available from: <http://dx.doi.org/10.1016/j.wneu.2011.08.033>
- Ke G, Meng Q, Wang T, Chen W, Ma W, Liu T-Y, Finley T, Wang T, Chen W, Ma W,

et al. 2017. LightGBM: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* [Internet].:3149–3157. Available from:

<http://papers.nips.cc/paper/6907-a-highly-efficient-gradient-boosting-decision-tree.pdf>

Kleinloog R, Mul N De, Post JA, Rinkel GJE. 2017. Risk Factors for Intracranial Aneurysm Rupture: A Systematic Review. *Neurosurgery* [Internet]. 82:431–440.

Available from: http://fdslive.oup.com/www.oup.com/pdf/production_in_progress.pdf

Larrabide I, Omedas P, Martelli Y, Planes X, Nieber M, Moya JA, Butakoff C, Sebastián R, Camara O, De Craene M, et al. 2009. GIMIAS: An open source framework for efficient development of research tools and clinical prototypes. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 5528:417–426.

Lauric A, Miller EL, Baharoglu MI, Malek AM. 2011. 3D shape analysis of intracranial aneurysms using the writhe number as a discriminant for rupture. *Ann Biomed Eng*. 39:1457–1469.

Lindgren AE, Koivisto T, Björkman J, von und zu Fraunberg M, Helin K, Jääskeläinen JE, Frösen J. 2016. Irregular Shape of Intracranial Aneurysm Indicates Rupture Risk Irrespective of Size in a Population-Based Cohort. *Stroke*.

Lorensen WE, Cline HE. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Comput Graph* [Internet]. [cited 2020 Jan 24]; 21:163–169. Available from: <http://portal.acm.org/citation.cfm?doid=37402.37422>

Millán RD, Dempere-Marco L, Pozo JM, Cebal JR, Frangi AF. 2007. Morphological characterization of intracranial aneurysms using 3-D moment invariants. *Med Imaging, IEEE Trans* [Internet]. 26:1270–1282. Available from:

<http://www.ncbi.nlm.nih.gov/pubmed/17896598>

Morel S, Diagbouga MR, Dupuy N, Sutter E, Braunersreuther V, Pelli G, Corniola M, Gondar R, Jägersberg M, Isidor N, et al. 2018. Correlating Clinical Risk Factors and Histological Features in Ruptured and Unruptured Human Intracranial Aneurysms: The Swiss AneuX Study. *J Neuropathol Exp Neurol* [Internet]. 77:555–566. Available from: <https://academic.oup.com/jnen/article/77/7/555/4982762>

Nieuwkamp DJ, Setz LE, Algra A, Linn FH, de Rooij NK, Rinkel GJ. 2009. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex,

and region: a meta-analysis. *Lancet Neurol* [Internet]. 8:635–642. Available from: [http://dx.doi.org/10.1016/S1474-4422\(09\)70126-7](http://dx.doi.org/10.1016/S1474-4422(09)70126-7)

Novotni M, Klein R. 2003. 3D Zernike Descriptors for Content Based Shape Retrieval. *Proc eighth ACM Symp Solid Model Appl* [Internet].:216–225. Available from: <http://portal.acm.org/citation.cfm?doid=781606.781639>

Raghavan ML, Ma B, Harbaugh RE. 2005. Quantified aneurysm shape and rupture risk. *J Neurosurg* [Internet]. 102:355–62. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15739566>

Suh SH, Cloft HJ, Huston J, Han KH, Kallmes DF. 2014. Interobserver variability of aneurysm morphology: Discrimination of the daughter sac. *J Neurointerv Surg* [Internet]. [cited 2019 Jan 23]; 8:38–41. Available from: <http://jn.is.bmj.com/>

UCAS Japan Investigators. 2012. The natural course of unruptured cerebral aneurysms in a Japanese cohort. *N Engl J Med* [Internet]. [cited 2019 Jan 23]; 366:2474–82. Available from: <https://www.nejm.org/doi/pdf/10.1056/NEJMoal113260>

Ujiie H, Tachi H, Hiramatsu O, Hazel AL, Matsumoto T, Ogasawara Y, Nakajima H, Hori T, Takakura K, Kajiya F. 1999. Effects of Size and Shape (Aspect Ratio) on the Hemodynamics of Saccular Aneurysms: A Possible Index for Surgical Treatment of Intracranial Aneurysms. *Neurosurgery* [Internet]. [cited 2019 Jan 23]; 45:119–130. Available from: <https://academic.oup.com/neurosurgery/article-lookup/doi/10.1097/00006123-199907000-00028>

Vlak MHM, Algra A, Brandenburg R, Rinkel GJE. 2011. Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: A systematic review and meta-analysis. *Lancet Neurol* [Internet]. 10:626–636. Available from: [http://dx.doi.org/10.1016/S1474-4422\(11\)70109-0](http://dx.doi.org/10.1016/S1474-4422(11)70109-0)

Wermer MJH, Van Der Schaaf IC, Algra A, Rinkel GJE. 2007. Risk of rupture of unruptured intracranial aneurysms in relation to patient and aneurysm characteristics: An updated meta-analysis. *Stroke*. 38:1404–1410.

Type	Sub-type	Details	#indices
Geometry Indices	Size indices	<ul style="list-style-type: none"> • Dome volume • Dome surface area • Neck diameter • Max. diameter • Height • Aneurysm size 	6
	Shape indices	<ul style="list-style-type: none"> • Non-sphericity index • Ellipticity index • Undulation index • Aspect ratio • Conicity factor • Bottleneck factor 	6
Distribution-derived features	Curvature metrics	<ul style="list-style-type: none"> • Gaussian and mean curvature • Distribution characteristics • Total curvature, normalized by surface area 	22
	Writhe metrics	<ul style="list-style-type: none"> • Free writhe and normalized inner-squared writhe • Distribution characteristics 	24
Moment-based descriptors	Zernike Moment Invariants (ZMI)	<ul style="list-style-type: none"> • Surface-based • Order $n = 20$ 	121
Total			179

Table 1. Composition of the feature pool for the morphological assessment of IAs.
#indices indicates the number of indices that a particular type contributes to the pool.

Attribute	Descriptions
<i>Rough surface</i>	Does the surface show an overall rough, non-smooth surface? Does it show structures that do not qualify as blebs or lobules?
<i>Blebs</i>	Are any blebs visible? A bleb is any localizable elevation of the dome surface whose volume is <i>smaller than 25%</i> of the primary dome compartment.
<i>Lobules</i>	Are any lobules visible? A lobule is any localizable elevation of the dome surface whose volume is <i>larger than 25%</i> of the primary dome compartment.
<i>Asymmetry</i>	Does the aneurysm appear asymmetric? Geometric asymmetry applies if the aneurysm dome lacks axes of symmetry.
<i>Complex vasculature</i>	Does the surrounding vasculature look complex such that it affects the overall perceived complexity of the aneurysm?
<i>Nothing applies</i>	None of the options above apply.

Table 2. Descriptions of the morphological attributes used in this study.

Model	Regressor	Feature space configuration			Motivation
		Selection	d	Repr.	
Ref.	SVRlin	Best univariate feature	1	ranked/ metric	Reference model using the best performing univariate feature of the pool.
A1	SVRlin	\mathcal{F}_{univ}	19	ranked	Combine statistically independent predictors with good univariate prediction in a multivariate model.
A2	SVRlin	$\mathcal{F}_{univ} \cup \mathcal{F}_{imp}$	31	ranked	
B	SVRlin	PCA, 90% of total variance	6	ranked	Reduce problem complexity by reducing redundancy in the data space. This assumes an (approximately) linear relationship.
C	GBM	\mathcal{F}_{imp}	31	metric	A nonlinear regression model may capture complex relationships between explanatory and predicted variables more accurately.

Table 3. Overview of the model configurations used in this study. d represents the number of dimensions of the reduced feature space; \mathcal{F}_{univ} and \mathcal{F}_{imp} are the set of features with the best univariate and most important candidates, respectively. SVRlin: support vector regression with linear kernel. GBM: gradient boosting machine. PCA: principal component analysis. Ref.: reference model. Repr.: data representation.

Morphological attribute	Hard comparison		Soft comparison	
	Fleiss' kappa κ_F		Cohen's kappa κ_C (mean \pm std)	Threshold τ_{jk}^* (mean \pm std)
Asymmetry	0.173	(slight-fair)	0.535 \pm 0.148	(moderate-substantial) 0.464 \pm 0.170
Rough surface	0.316	(fair-moderate)	0.659 \pm 0.094	(substantial) 0.397 \pm 0.187
Blebs	0.274	(fair)	0.625 \pm 0.075	(substantial) 0.453 \pm 0.191
Lobules	0.282	(fair-moderate)	0.647 \pm 0.117	(substantial) 0.438 \pm 0.223
Complex parent vasculature	0.175	(slight-fair)	0.523 \pm 0.143	(moderate) 0.322 \pm 0.171

Table 4. Average agreement for the binary ratings on the morphological attributes, evaluated using hard and soft comparisons of raters (see text). The data comprises ratings for 134 cases from 26 different raters (16 instructed laypersons, 10 clinical experts). Our results suggest that the raters substantially agree if the rater subjectivity is taken into account, and that agreement varies across different attributes.

Predicted variable	Data	Model	RMSE			R ²		
			<i>mean</i>	<i>std</i>	<i>p-val</i>	<i>mean</i>	<i>std</i>	<i>p-val</i>
Perceived irregularity	Ranked	<i>Reference</i>	0.129	0.016	Ref.	0.788	0.067	Ref.
		<i>Model A1</i>	0.122	0.016	< 0.001	0.809	0.067	< 0.001
		<i>Model A2</i>	0.113	0.015	< 0.001	0.836	0.051	< 0.001
		<i>Model B</i>	0.129	0.016	< 0.001	0.786	0.068	< 0.001
	Metric	<i>Reference</i>	0.150	0.018	Ref.	0.677	0.085	Ref.
		<i>Model C</i>	0.109	0.012	< 0.001	0.829	0.055	< 0.001
Rough surface	Ranked	<i>Reference</i>	0.228	0.026	Ref.	0.464	0.138	Ref.
		<i>Model A2</i>	0.216	0.027	< 0.001	0.513	0.144	< 0.001
Blebs	Ranked	<i>Reference</i>	0.203	0.023	Ref.	0.511	0.133	Ref.
		<i>Model A2</i>	0.189	0.024	< 0.001	0.577	0.125	< 0.001
Lobules	Ranked	<i>Reference</i>	0.203	0.037	Ref.	0.510	0.170	Ref.
		<i>Model A2</i>	0.174	0.024	< 0.001	0.638	0.109	< 0.001
Asymmetry	Ranked	<i>Reference</i>	0.202	0.022	Ref.	0.492	0.141	Ref.
		<i>Model A2</i>	0.172	0.020	< 0.001	0.627	0.114	< 0.001
Complex vasc.	Ranked	<i>Reference</i>	0.300	0.027	Ref.	0.032	0.137	Ref.
			0.293	0.027	> 0.05	0.070	0.179	> 0.05

Table 5. Summary of the prediction performances for the different multivariate model configurations used to predict the perceived irregularity (upper half) and the morphological attributes (lower half). The models were trained and validated in a nested cross-validation scheme with 50 repetitions. For perceived irregularity, the best performing *univariate* model (based on the curvature metric GLN) is given as reference. We evaluated the models for ranked and non-ranked data representation, where both explanatory and predicted variables were ranked prior to training. Root mean squared error (RMSE) and the coefficient of determination (R^2) are provided. For the morphological attributes, we report the results of the best-performing univariate and multivariate models.

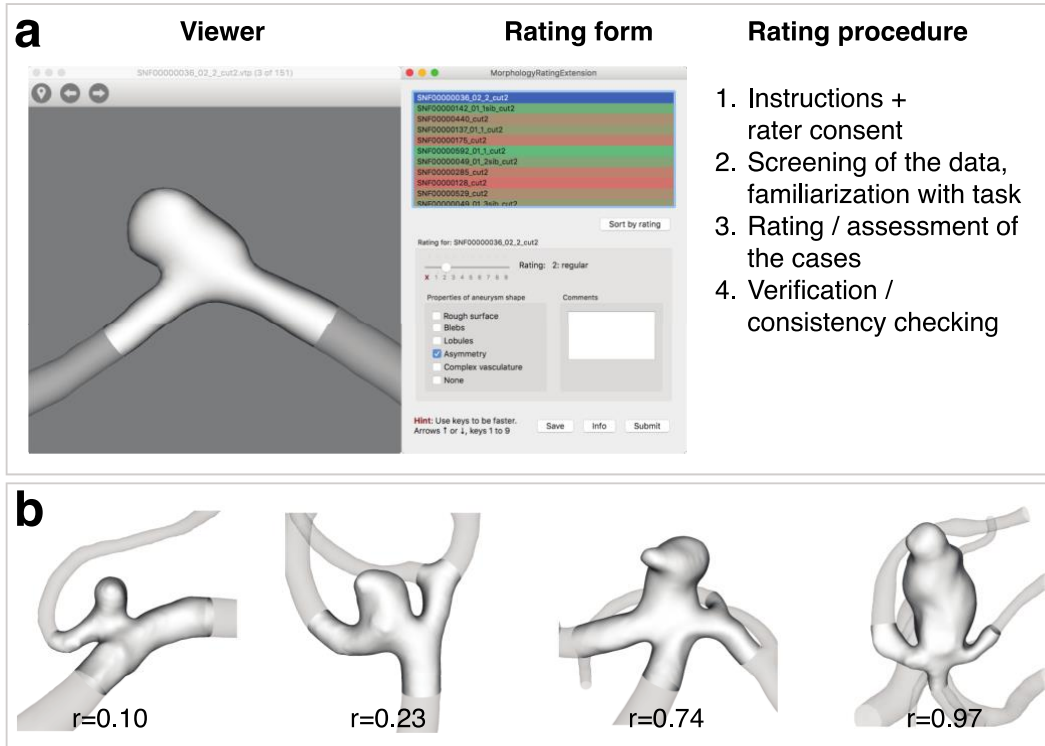


Figure 1. (a) Screenshot of the rating tool for interactive display of 3D geometries and rating acquisition, here for IA morphology: perceived irregularity (ordinal assessment) and a list of morphological attributes (binary assessments). The tool facilitates the efficient comparison of cases and rating verification. (b) Exemplary IA geometries ordered by increasing perceived irregularity from very regular ($r'_i = 0$) to very irregular ($r'_i = 1$).

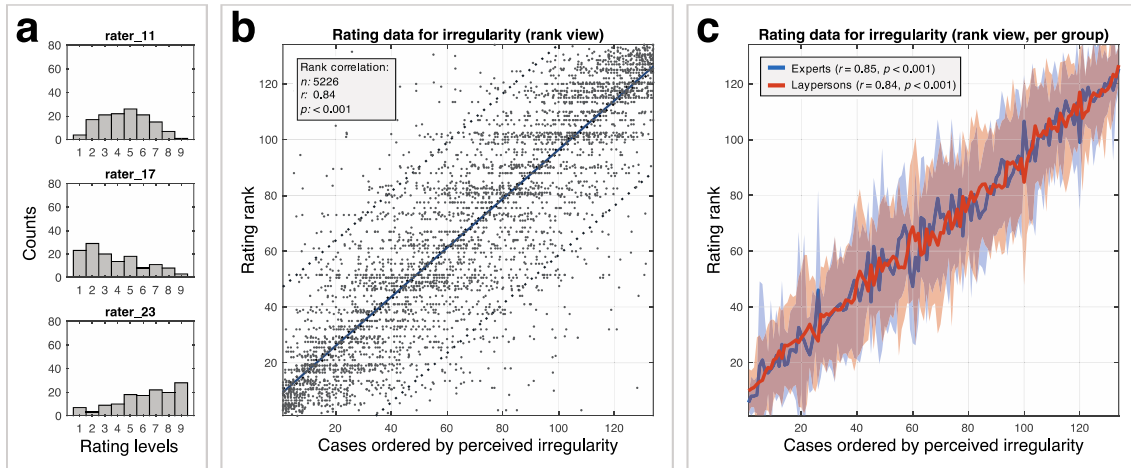


Figure 2. (a) Exemplary histograms summarizing the ordinal irregularity ratings of three different raters, demonstrating different rating biases. (b) Scatter plot showing the ratings by the 39 included raters for the 134 aneurysms ($n = 5226$ data points, ranked per rater, Spearman rank correlation $\rho_{Sp} = 0.84$ ($p < 0.001$) between the individual rating ranks and the aggregated). The plot also shows the regression line and its 95% tolerance- and confidence intervals (dotted lines). (c) Data stratified by rater sub-cohort (clinical experts vs. instructed laypersons). Solid lines: mean rating ranks per aneurysm. Shaded areas: \pm standard deviation of rating ranks.

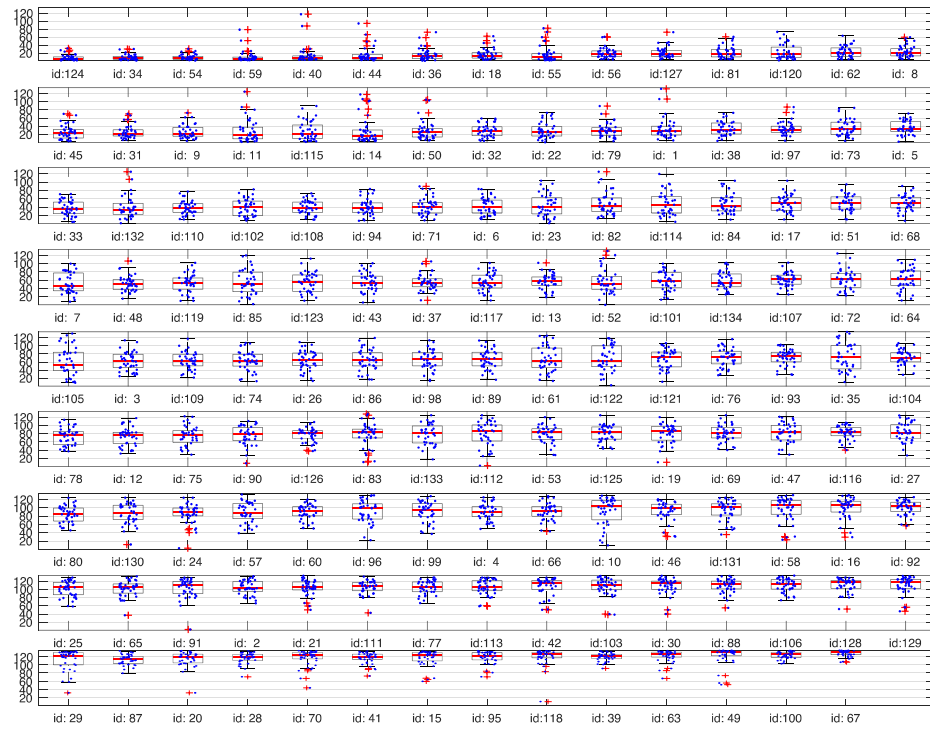


Figure 3. The irregularity rating ranks by all 39 raters for all 134 cases. The cases are sorted by increasing mean. By comparing the data spread, one can observe that the inter-rater agreement varies considerably between different cases. As a trend, the agreement is high for the extreme cases (very regular, very irregular).

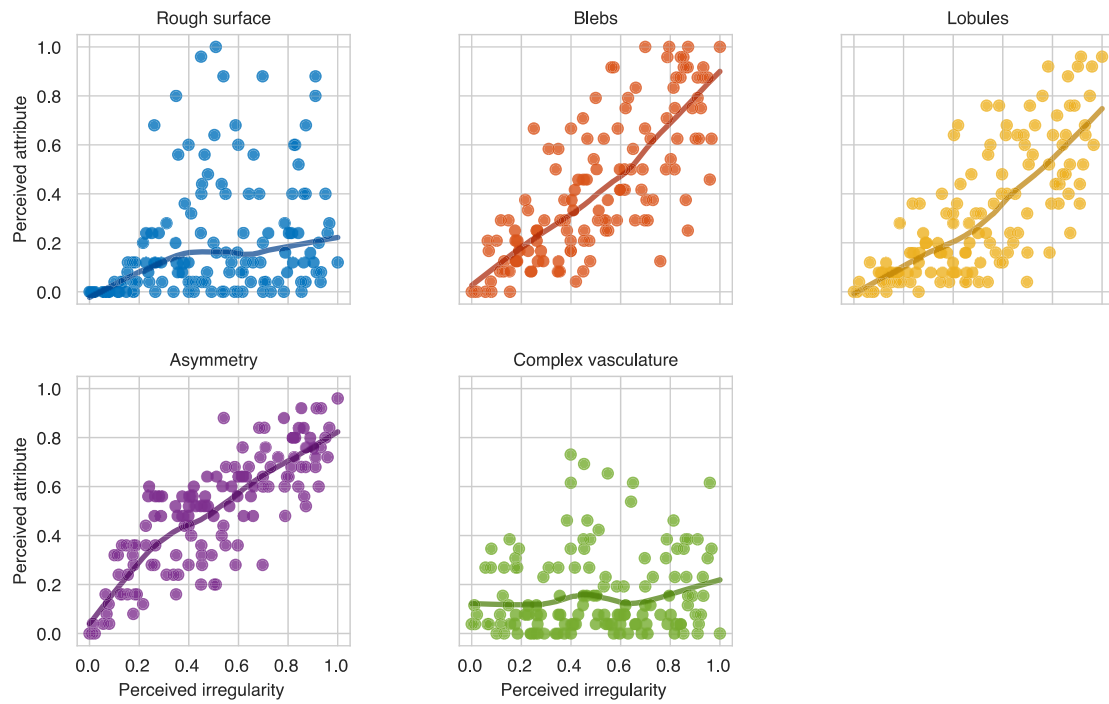


Figure 4. Aggregated ratings for the morphological attributes. The plots show the data (dots) for the 134 cases, comparing the perceived irregularity (abscissa) with the aggregated ratings (relative counts) of the following six attributes (multiple choices allowed): asymmetry, rough surface, blebs, lobules, complex parent-vasculature and nothing (if none of the characteristics applied). We also show LOWESS regression curves (with smoothing factor 0.2) to identify possible trends in the ratings.

Online Supplemental Material

Predictor	Correlation ρ_{sp}	
	A=0.01 mm ²	A=0.05 mm ²
GI: curvature (Gaussian, L ₂ N)	0.89	0.92
GI: curvature (Gaussian, stdN)	0.89	0.92
GI: curvature (mean, L ₂ N)	0.88	0.91
GI: curvature (mean, vL ₁)	0.87	0.88
GI: curvature (mean, stdN)	0.85	0.89
writhe: inner squared (H)	0.84	0.83
writhe: inner squared (mean)	0.84	0.83
GI: curvature (Gaussian, L ₂ N)	0.84	0.84
GI: shape (NSI)	0.80	0.80
writhe: inner squared (std)	0.79	0.78
writhe: inner squared (μ_2)	0.77	0.77
GI: shape (EI)	0.76	0.76
ZMI: cumulant (n40)	0.74	0.73
ZMI: cumulant (n10)	0.74	0.74
ZMI: cumulant (n20)	0.74	0.73
GI: curvature (Gaussian, L ₂ NCH)	0.73	0.77
ZMI: cumulant (n05)	0.73	0.73
GI: shape (BF)	0.72	0.72
GI: size (aSz)	0.70	0.70
GI: shape (UI)	0.67	0.71
GI: curvature (mean, L ₂ NCH)	0.66	0.76

Table A. Best performing univariate predictors for perceived irregularity, evaluated for two different average mesh cell areas A=0.01mm² and A=0.05mm². We included only metrics with Spearman correlation $\rho_{sp} > 0.7$. The overall ordering of the features appears relatively stable for the two different mesh sizes examined. Only curvature metrics yielded systematically higher coefficients. All metrics have been computed on 3D geometries of the aneurysm dome. Their implementation follows the references cited in the main article. Abbreviations: Curvature L₂N – total curvature (L₂-norm), normalized by the surface area; curvature stdN – standard deviation of curvature, normalized by surface area; curvature vL₁ – area weighted variance of the curvature; curvature L₂NCH – same as curvature L₂N but further normalized by the total curvature (L₂N) of the convex hull; writhe mean, std, H, μ_2 : mean, standard deviation, entropy or second statistical moment of the writhe values for a surface; GI – geometry indices; NSI – non-sphericity index; EI – ellipticity index; UI – undulation index; BF – bottleneck

factor; aSz – aneurysm size; ZMI cumulant – metrics derived from Zernike moment invariants.

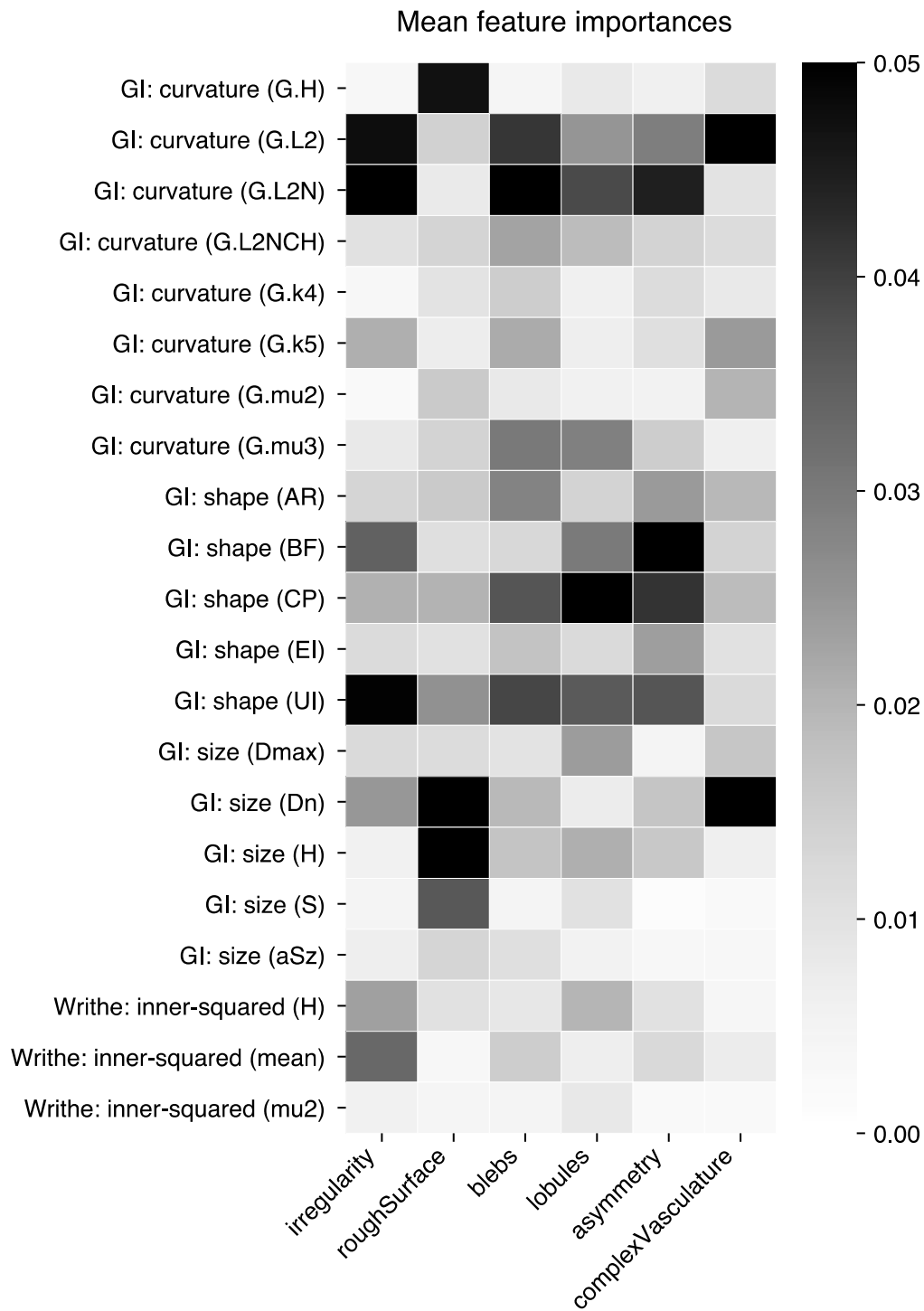


Figure A. Comparison of the mean feature importance (FI) for the prediction of perceived characteristics, averaged over the 1000 gradient boosting machines (GBMs) trained in the feature selection step. FI measures how valuable a feature was when training GBMs. Black and white colour indicate high and low FI, respectively. The listing is freed from highly redundant features and features that show low importance in all morphological characteristics. Abbreviations: see caption of Table A.